

Using Data Linkage Software: When Two Heads Are Better Than One

Lisa Wyman, MPH
Bureau of Epidemiology

Kathy Bell, DVM
University of Utah



Types of Linkage

- One-to-one relationship
- One-to-many relationship
- Unduplication

One-to-One Relationship

- Two sets of data are compared
- Goal is to identify all the “best” pairs or matches between the sets
- Example: Mortality linkages
 - Only possible for one death record to match with influenza-associated hospitalization data

One-to-Many Relationship

- Two sets of data are compared
- Goal is to identify all elements of one set that match to a particular element of a second
- Example: Geocoding linkages
 - Possible for any specific address can match to multiple entries in a cancer registry file
 - Multiple records within one family occur

Unduplication

- Single data set used
- Goal is to identify multiple matches within a single set of data
- Example: NETSS unduplication
 - Removes multiple entries for same individual and disease event

Computer Based Linkages

- Vary widely
 - Simple home grown code modules
 - Complex custom written stand alone programs
 - Entire software suites
- Modern algorithms generally fall into two classes:
 - Deterministic
 - Probabilistic

Deterministic Linkages

- Depend upon an entity relationship between the data elements being compared
 - Static, predefined, and empirically based
- Rules used to match the data are the same
 - Regardless of data file size, missing values, what values are present in data
- Linkage software example: Link King

Probabilistic Linkages

- Utilize a more fluid relationship entity relationship
- Take into account various attributes of the data important for the increasing probability of match situations
 - Weights assigned to individual fields
- Examples of incorporated attributes:
 - Error rates
 - Frequency analysis of values
- Linkage software example: LinkSolv

Link King: Linking Influenza Hospitalizations and Death Records

Influenza-Associated Hospitalizations

- Reportable condition in Utah
 - Defined as laboratory confirmation of influenza and hospital admission
 - Not to be confused with hospital discharge data (code 487)
- 1032 hospitalizations reported total for 2004-05, 2005-06, and 2006-07 influenza seasons

Pneumonia and Influenza Deaths

- Identified via EDEN
 - Free text field key word search
 - Primary cause of death
- 735 deaths identified for the 2006-07 season

Linking Data Sets

- Purposes of project:
 - Identify those hospitalized cases that died
 - Characterize those cases
 - Demographics
 - Clinical aspects
 - Become familiar with linkage software
 - Link King

Link King: Overview

- Primarily deterministic linkage software
 - Some probabilistic features
- FREE!
- Available on the Internet
 - www.the-link-king.com
- Adapted primarily from Washington State's Division of Alcohol and Substance Abuse
- Good for both unduplication and record linkage
 - SAS based

Link King: What's Needed

- Required:
 - First name
 - Last name
 - Date of birth or SSN
- Recommended:
 - Middle name
 - Maiden name
 - Gender
 - Race/ethnicity
 - Zip code

Dedupe Software / Record Linkage Software by The Link King FREE ! - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Reload Print Mail News RSS Feeds

Address <http://www.the-link-king.com/> Go Links »

Home Screens/Video Stats #1 Stats #2 Contact Us Download Unix/Linux Info Royal Links

The Link King

Record Linkage and Consolidation Software

In the realm of public domain software for record linkage and unduplication (aka. **dedupe software**), The Link King reigns supreme. The Link King has fashioned a powerful alliance between sophisticated probabilistic record linkage and deterministic record linkage protocols incorporating features unavailable in many proprietary record linkage programs. ([detailed overview \(pdf\)](#))

The Link King's probabilistic record linkage protocol was adapted from the algorithm developed by MEDSTAT for the Substance Abuse and Mental Health Services Administration's (SAMHSA) Integrated Database Project. The deterministic record linkage protocols were developed at Washington State's Division of Alcohol and Substance Abuse for use in a variety of evaluation and research projects.

The Link King's graphical user interface (GUI) makes record linkage and unduplication easy for beginning and advanced users. The data linking neophyte will appreciate the easy-to-follow instructions. The Link King's artificial intelligence will assist in the selection of the most appropriate linkage/unduplication protocol. The technical wizard will appreciate the discussion of data linkage/unduplication issues in The Link King's user manual, the variety of user-specified options for blocking and linkage decisions, and the powerful interface for manual review of "uncertain" linkages.


The Link King requires a base SAS license but NO SAS programming experience. *If you have a SAS dataset, a SPSS portable file, an EXCEL spreadsheet, or a comma/tab delimited file, The Link King can do the rest.* The Link King prefers SAS v9.0 or higher but is comfortable with SAS v8.

Best of all, [download](#) The Link King for FREE !

Data Elements Used in Linking

Required:

Optional:



[Logo by Reid Psaltis](#)

The Link King v6.3.4

with AutoUpdate
Last updated 2/29/08

Now Playing !

8-minute Link King Demo Video
[View Online \(flash video\)](#)
[Download \(for Windows Media Player\)](#)

Link King Screensaver

[Preview Screensaver \(flash video\)](#)

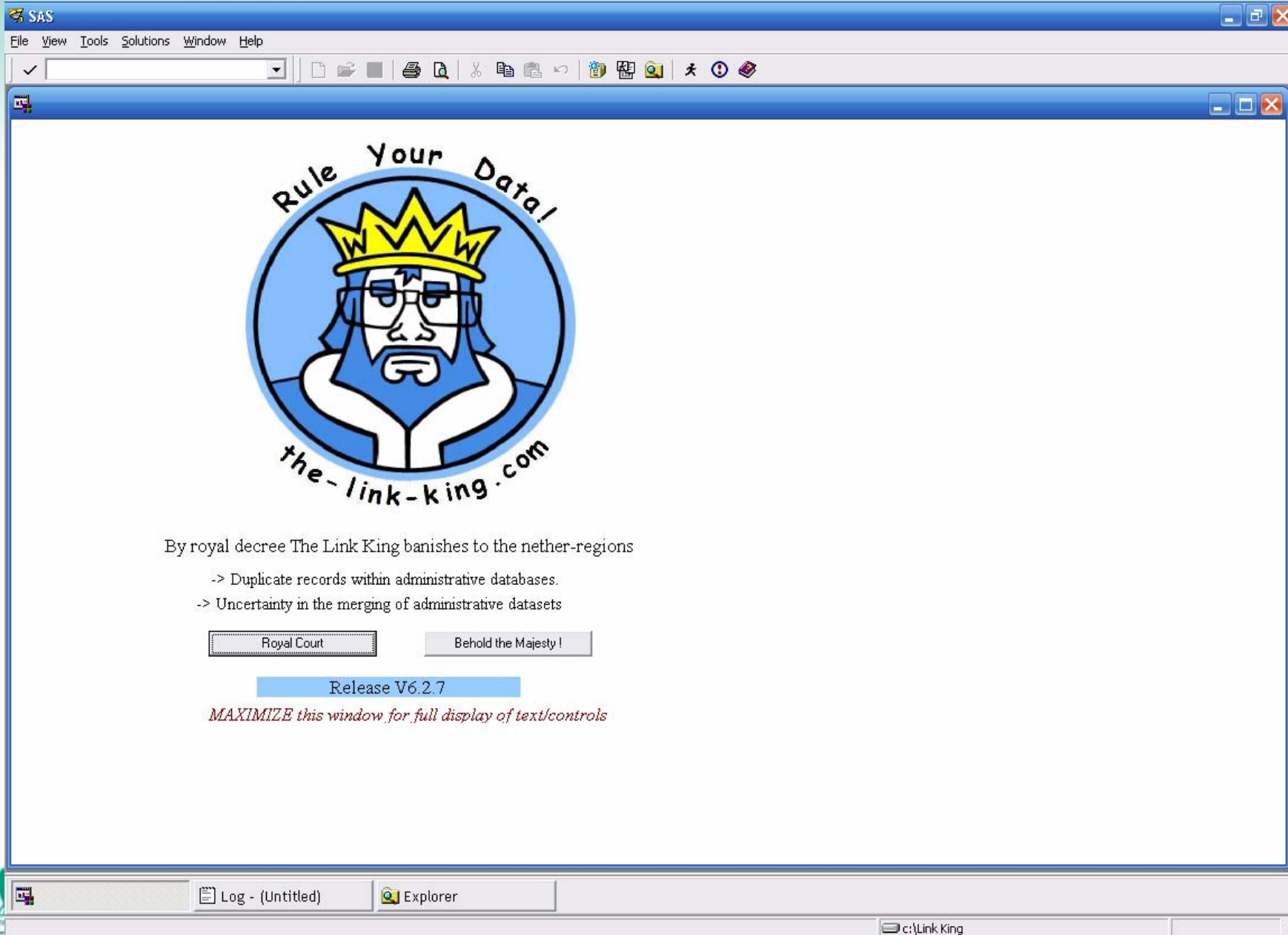
Empirical Evaluations of Link King Accuracy

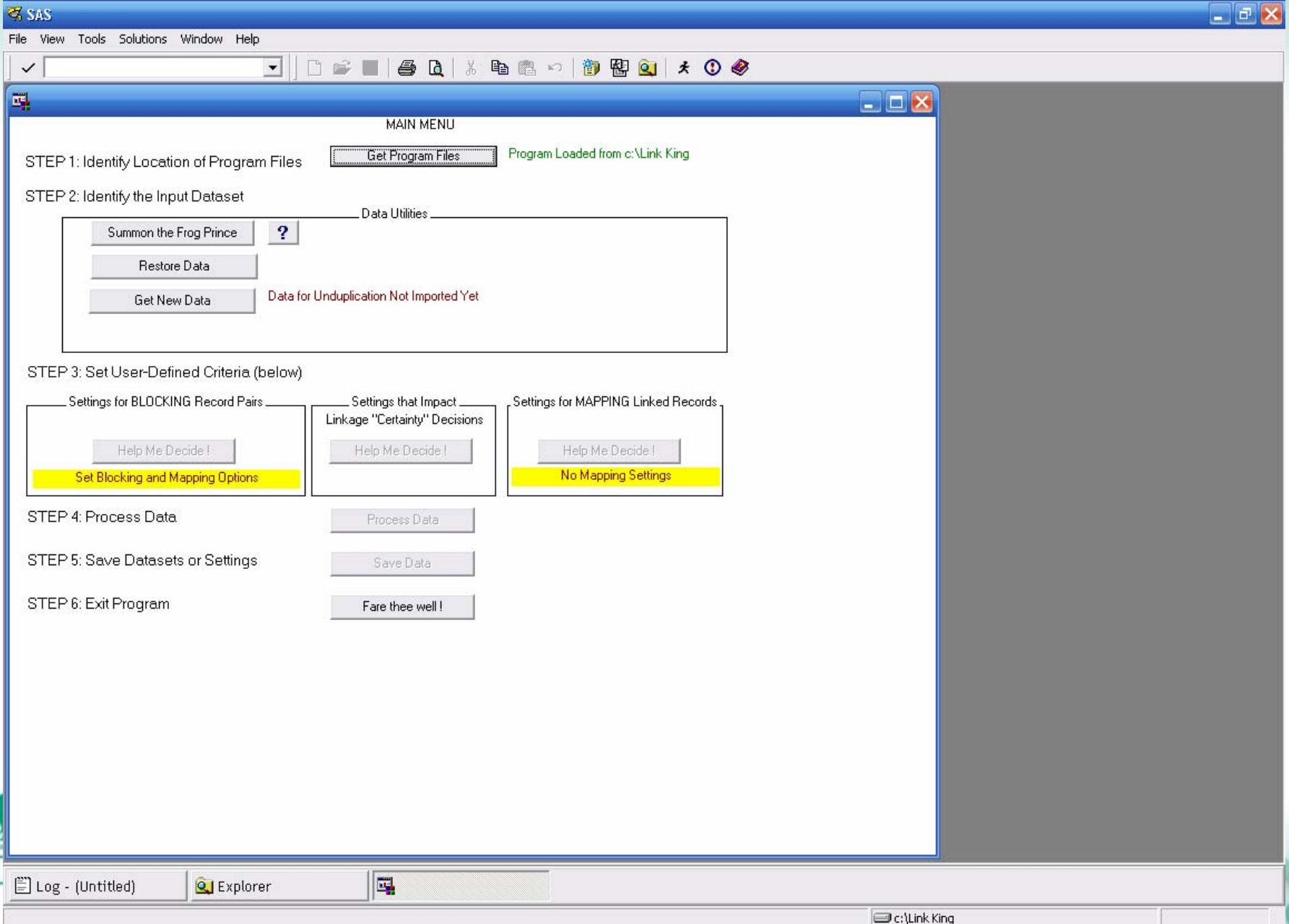
[Health Informatics Journal](#)
[Kansas Health Institute](#)

Features:

- Probabilistic Record Linkage Algorithm
- Deterministic Record Linkage Algorithm
- Calculates distance between zip code centroids for use in probabilistic algorithm.

Done Internet





Link King Steps

- Create SAS datasets
- Unduplicate datasets
- Determine variables to match on
 - Last name
 - First name
 - Date of birth
 - Gender
 - Race
 - Ethnicity
- Review matches



Import Data for Linking/Unduplication

Step 1: Select SAMPLE Dataset

Data Format: SAS

HSP4567

Client Identifier: ID Prefix (option)

First Name: FIRST

Middle Name:

Last Name: LASTNAME

Maiden Name:

Social Security Number:

Birthdate: BIRTHDATE

Gender: SEX

Race / Ethnicity: ETHNIC

"Flex" Variable: ZIPCODE

"Flex" Vars

Import Browse

Step 2: Select MATCHING Dataset (if applicable)

Data Format: SAS

PIDEATH67

Client Identifier: Prefix (option)

First Name: FIRST

Middle Name: MIDDLE

Last Name: LASTNAME

Maiden Name:

Social Security Number:

Birthdate: DOB

Gender: SEX

Race / Ethnicity: ETHNIC

"Flex" Variable:

Import Browse

Navigate using

Matching Protocol for FLEX

Results

- Seven linkages
 - Only 4 marked as having died in hospitalization database
- Primarily hospitalizations from 2006-07 season (as expected)
 - 2 hospitalized during 2005-06 season
- 6 persons ≥ 65 years, 1 pediatric case

Manual Data Linkage

- Pneumonia and influenza death certificate and hospitalization data sorted by
 - Last name
 - Date of birth

Matches

- The same 7 people were matched in both data sets manually and by Link King
- 3 of the 7 people were not marked as died in the hospitalization data yet had a death certificate

Lessons Learned

- Link King program matching was accurate as verified manually
- Manual linking of very large datasets would be difficult and time consuming
- Deterministic linkages limiting in requiring names and date of birth/ SSN

Future Projects

- Linking deaths from 2007-08 with hospitalizations
 - Same process used
- Linking influenza-associated hospitalizations with hospital discharge data
 - Major undertaking
 - Discharge data missing name in most cases
 - SSN and MRN missing for most IAHs

Discharge Data Linkage

- Deterministic linkage not possible for whole dataset at this point
 - Limited to those cases with known name
 - May be still worthwhile
 - Done manually for the 2004-05 season
 - Gave insight into causes of hospitalizations for IAHs
 - No necessarily influenza

Acknowledgments

- Jeff Duncan
- Wu Xu
- University of Southern California
- David Jackson